Distribution - Free Continuous Bayesian Belief Nets

D. Kurowicka
Delft Institute for Applied Mathematics
Delft Univeristy of Technology
Delft
The Netherlands
D.Kurowicka@ewi.tudelft.nl

R.M.Cooke
Delft Institute for Applied Mathematics
Delft Univeristy of Technology
Delft
The Netherlands
R.M.Cooke@ewi.tudelft.nl

Abstract

This paper introduces distribution free continuous belief nets using the vine - copulae modelling approach. Nodes are associated with arbitrary continuous invrtible distributions, influences are associated with (conditional) rank correlations and are realized by (conditional) copulae. Any copula which represents (conditional) independence as zero (conditional) correlation can be used. We illustrate this approach with a flight crew alertness model.

1 Introduction

Bayesian belief nets (bbns) are directed acyclic graphs representing high dimensional uncertainty distributions, and are becoming increasingly popular in modelling complex systems. Application is severely limited by the excessive assessment burden, which leads to informal and indefensible quantification. Continuous bbns exist only for the joint normal case, and generally require assessing entire covariance matrices.

In (Kurowicka D. and Cooke R.M. 2002) the authors introduced an approach to continuous bbns using vines (Bedford T.J., Cooke R.M. 2002) and the elliptical copula (Kurowicka D, Misiewicz J. and Cooke R.M. 2000). Influences were associated with conditional rank correlations, and these were realized by (conditional) elliptical copulae. While this approach has some attractive features, notably in preserving some relations between conditional and partial correlation, it also has disadvantages. Foremost among these is the fact that zero (conditional) correlation does not correspond to (conditional) independence under the elliptical copula. This motivates the "copula-free" approach advanced here; any copula may be used so long as the chosen copula represents (conditional) independence as zero (conditional) correlation. This approach cannot rely on the equality of partial and conditional correlation, and hence cannot rely on vine transformations to deal with observation and updating, as done in (Kurowicka D. and Cooke R.M. 2002).
Nonetheless, it is shown that the elicitation protocol of (Kurowicka D. and Cooke R.M. 2002) based on conditional rank correlation can work in a copula-free environment. A unique joint distribution can be determined and sampled based on the protocol, which factorizes in the manner prescribed by the bbn. Further, this distribution can be updated with observations. We note that quantifying bbn's in this way requires assessing *all* (continuous, invertible) one dimensional marginal distributions. On the other hand, the dependence structure is meaningful for *any* such quantification. In fact, when comparing different decisions or assessing the value of different observations, it is frequently sufficient to observe the effects on the quantile functions of each node. For such comparisons we do not need to assess the one dimensional margins at all. This will be illustrated in the sequel.

We assume the reader is familiar with bbn's. We introduce vines and copula's in section 1, and distribution free bbn's in section 3. Section 4 illustrates their use with an example from (Roelen A.L.C. et. al. 2003) involving airline flight crew alertness.

2 Vines and copulae

A graphical model called vines was introduced in (Cooke R.M. 1997), (Bedford T.J., Cooke R.M. 2002). A vine on N variables is a nested set of trees, where the edges of tree j are the nodes of tree j+1, and each tree has the maximum number of edges. A regular vine on N variables is a vine in which two edges in tree j are joined by an edge in tree j+1 only if these edges share a common node. A regular vine is called a canonical vine if each tree T_i has a unique node of degree N-i, hence has maximum degree. A regular vine is called a D-vine if all nodes in T_1 have degree not higher than 2 (see Figure 1). There are N(N-1)/2 edges in a regular vine on N variables. Each edge in a regular vine may be associated with a constant conditional rank correlation (for j=1 the conditions are vacuous) and, using a copula (bivariate distribution on a unit square with uniform margins), a joint distribution satisfying the vine-copula specification can be constructed and sampled on the fly (Cooke R.M. 1997). The conditional rank correlations associated with each edge are determined as follows: the variables reachable from a given edge are called the constraint set of that edge. When two edges are joined by an edge of the next tree, the intersection of the respective constraint sets are the conditioning variables, and the symmetric differences of the constraint sets are the conditioned variables. The regularity condition insures that the symmetric difference of the constraint sets always contains two variables. Each pair of variables occurs once as conditioned variables. For the precise definitions and all properties of a regular vine we refer to (Bedford T.J., Cooke R.M. 2002).

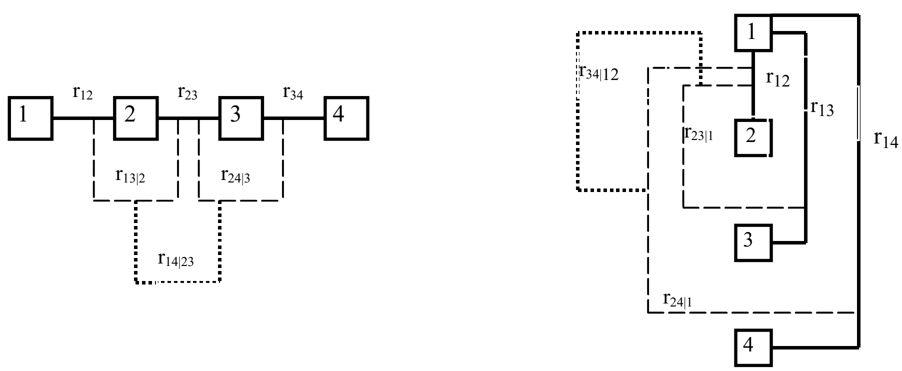


Figure 1: D-vine (left) and canonical vine (right) on 4 variables with (conditional) rank correlations assigned to the edges.

The rank correlation specification on regular vine plus copula determines the whole joint distribution. To sample a distribution specified by the D-vine in Figure 1, D(1,2,3,4) the following algorithm can be used: Sample four independent variables distributed uniformly on interval [0,1], U_1, U_2, U_3, U_4 and calculate values of correlated variables X_1, X_2, X_3, X_4 as follows
Sampling procedure for the D-vine in Figure 1 is the following

1. $x_1 = u_1$,
2. $x_2 = F^{-1}_{r_{12};x_1}(u_2)$,
3. $x_3 = F^{-1}_{r_{23};x_2}\left(F^{-1}_{r_{13|2};F_{r_{12};x_2}(x_1)}(u_3)\right)$,
4. $x_4 = F^{-1}_{r_{34};x_3}\left(F^{-1}_{r_{24|3};F_{r_{23};x_3}(x_2)}\left(F^{-1}_{r_{14|23};F_{r_{13|2};F_{r_{23};x_2}(x_3)}(F_{r_{12};x_2}(x_1))}(u_4)\right)\right)$

where where $F_{r_{ij|k};X_i}(X_j)$ denotes the cumulative distribution function for X_j, applied to X_j, given X_i under the conditional copula with correlation $r_{ij|k}$.

Notice that the sampling procedure for D-vine uses conditional distributions as well as inverse conditional distributions. We shorten the notation by dropping the "r"'s and write the general sampling algorithm as:

$$
\begin{aligned}
x_1 &= u_1, \\
x_2 &= F^{-1}_{2|1:x_1}(u_2), \\
x_3 &= F^{-1}_{3|2:x_2}\left(F^{-1}_{3|12:F_{1|2}(x_1)}(u_3)\right), \\
x_4 &= F^{-1}_{4|3:x_3}\left(F^{-1}_{4|23:F_{2|3}(x_2)}\left(F^{-1}_{4|123:F_{1|23}(x_1)}(u_4)\right)\right), \\
x_5 &= F^{-1}_{5|4:x_4}\left(F^{-1}_{5|34:F_{3|4}(x_3)}\left(F^{-1}_{5|234:F_{2|34}(x_2)}\left(F^{-1}_{5|1234:F_{1|234}(x_1)}(u_5)\right)\right)\right), \\
\dots \\
x_n &= F^{-1}_{n|n-1:x_{n-1}}\left(F^{-1}_{n|n-2,n-1:F_{n-2|n-1}(x_{n-2})}\left(F^{-1}_{n|n-3,n-2,n-1:F_{n-3|n-2,n-1}(x_{n-3})}\left(\dots\right.\right.\right. \\
&\qquad \left.\left.\left.\left(F^{-1}_{n|1\dots n-1:F_{1|2\dots n-1}(x_1)}(u_n)\right)\dots\right)\right)\right);
\end{aligned}
$$

We build complex high dimensional distributions from two dimensional and conditional two dimensional distributions with uniform margins. The two dimensional distributions are called "copulae".

Definition 2.1 [Copula] *A copula C is a distribution on the unit square with uniform margins.*

Definition 2.2 *Random variables X and Y are joined by copula C if their joint distribution can be written*

$$F_{XY}(x, y) = C(F_X(x), F_Y(x)).$$

The diagonal band copula was introduced in (Cooke R.M. and Waij R. 1986). For positive correlation the mass is concentrated on the diagonal band with vertical bandwidth $\beta = 1-\alpha$. Mass is distributed uniformly on the rectangle with corners $(\beta, 0), (0, \beta), (1-\beta, 1)$ and $(1, 1-\beta)$ and is uniform but "twice as thick" in the triangular corners. The correlation coefficient is given by $\rho = sign(\alpha)\left((1-|\alpha|)^3 - 2(1-|\alpha|)^2 + 1\right)$ (Cooke R.M. and Waij R. 1986). For negative correlation the band is drawn between the other corners. Figure 2 shows the density of the diagonal band distribution with correlation 0.8.

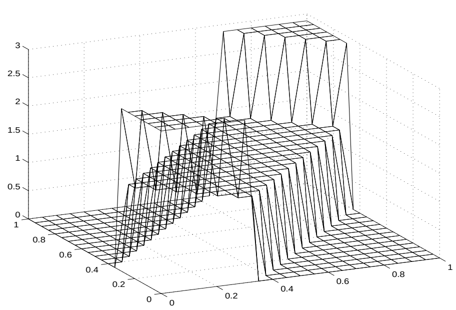


Figure 2: A density of the diagonal band copula with correlation 0.8.

In this paper we use the diagonal band copula because of its compliant analytical form . However we could use any copulae for which zero correlation entails independence, including the maximum entropy copulae, Frank's copulae, etc. For more information about copulae we refer to e.g. (Joe H. 1997), (Doruet Mari D., Kotz S. 2001), (Nelsen R.B. 1999).

3 Continuous BBNs

Let us associate nodes of a bbn with continuous univariate random variables and to interpret "influence" in terms of (conditional) rank correlations. We assume throughout that all univariate distributions have been transformed to uniform distributions on $(0, 1)$. To determine which (conditional) correlations are necessary we follow the following protocol:

1. Construct a (generally non-unique) sampling order for the nodes, that is, an ordering such that all ancestors of node i appear before i in the ordering. Index the nodes according to the sampling order $1, \dots, n$.

2. Factorize the joint in the standard way following the sampling order. Underscore those nodes in each condition which are not necessary in sampling the conditioned variable. If the sampling order is $1, 2, \dots, n,$, and if in sampling $2, \dots, n$ variable 1 is not necessary then we have:

$$P(1, \dots, n) \quad = \quad P(1)P(2|\underline{1})P(3|2\underline{1}) \dots P(n|n-1, n-2, \dots, \underline{1}).$$

   If we drop the underscored variables, we obtain the standard factorization for the bbn. Notice that since 1 is independent of $2, \dots, n$ then only the following distributions must be specified:

$$\{(32), (43), (42|3), \dots, (n, n-1), (n, n-2|n-1), \dots, (n, 2|n-1, \dots, 3)\}.$$

3. For each term K in the factorization $K = 1, \dots, n$ we define set of conditional variables $\mathcal{C}^K$ (non-underscored variables in term K of the factorization) and the set of independent variables $\mathcal{I}^K$ (underscored variables in term K). Let $\mathcal{C}^K = \{c_1^K, \dots, c_{j_K}^K\}$ and $\mathcal{I}^K = \{i_1^K, \dots, i_{s_K}^K\}$, where $i_t^K \neq c_h^K$, $s_K + j_K = K - 1$ (for $P(4|32\underline{1})$ $C^4 = \{3, 2\}$ and $I^4 = \{1\}$) then for K-th factorization the following correlations must be assessed

$$r_{21}, r_{3c_1^3}, r_{3c_2^3|c_1^3}, r_{3c_{j_3}^3|c_{j_3-1}^3,\dots,c_1^3}, \dots, r_{nc_1^n}, r_{nc_2^n|c_1^n}, \dots, r_{nc_{j_n}^n|c_{j_n-1}^n,\dots c_1^n} \tag{1}$$

To sample a distribution specified by a continuous bbn we use the sampling procedure for D-vine. For each term of the factorization we build a D-vine on K variables denoted by $\mathcal{D}^K = D(K, \mathcal{C}^K, \mathcal{I}^K)$. The ordering of variables is very important. We start with the K dependent variables and end with the independent variables. Notice that (conditional) correlations assigned to the left most branches of $\mathcal{D}^K$ are those in (1) and

$$r_{Ki_1^K|c_{j_K}^K,\dots c_1^K} = r_{Ki_2^K|i_1^K,c_{j_K}^K,\dots c_1^K} = \dots = r_{Ki_s^K|i_{s-1}^K,\dots,i_1^K,c_{j_K}^K,\dots c_1^K} = 0 \tag{2}$$

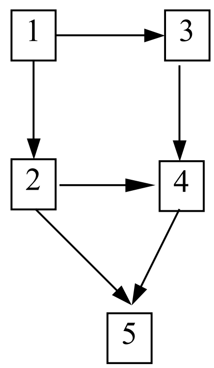
because variable K and the variables in $\mathcal{I}^K$ are independent given $\mathcal{C}^K$. We sample X_K using the sampling procedure for $\mathcal{D}^K$. In general it is not possible to keep the same order of variables in each $\mathcal{D}^K$. Hence in sampling procedure some conditional distributions will have to be calculated as in Example **??**. If the coupula represents (conditional)independence as zero (conditional)correlation, then this sampling routine realizes the factorization of the bbn.

Theorem 3.1 *Given a bbn, specification of (1), (2) and a copula for which correlation 0 entails independence uniquely determines the joint distribution.*

Proof. Each term of the factorization involves K variables $1, 2, \dots, K$. $\mathcal{D}^2$ with (conditional) rank correlations specified by (1) and (2) and a copula gives a bivariate distribution (1,2). This distribution and $\mathcal{D}^3$ with (conditional) rank correlations of the most left branches of the vine given by (1) and (2) determines the

distribution of (1,2,3) and so on. Finally $\mathcal{D}^n$ with correlations of the most left branches of the vine given by (1) and (2) and the n-1-dimensional distribution $(1,2,\ldots,n-1)$ suffice to construct $(1,2,\ldots,n)$. Uniqueness of this specification follows from uniqueness of rank correlation specification on a regular vine. □
The following example shows that this approach may require calculating multidimensional integrals.

Example 3.1 *Let us consider the following bbn on 5 variables.*



Sampling order: 1, 2, 3, 4, 5.

Factorization: $P(1)P(2|1)P(3|\underline{2}1)P(4|32\underline{1})P(5|4\underline{3}2\underline{1})$.

Rank correlations that have to be assessed:

$$r_{21}, r_{31}, r_{43}, r_{42|3}, r_{54}, r_{52|4}.$$

In this case $\mathcal{D}^4 = D(4,3,2,1)$ but the order of variables in $\mathcal{D}^5$ must be $D(5,4,2,3,1)$. Hence this bbn cannot be represented as one vine.

Using conditional independence properties of the bbn, the sampling procedure can be simplified as:

$$x_1 = u_1,$$

$$x_2 = F^{-1}_{r_{21};x_1}(u_2),$$

$$x_3 = F^{-1}_{r_{31};x_1}(u_3),$$

$$x_4 = F^{-1}_{r_{43};x_3}\left(F^{-1}_{r_{42|3};F_{2|3}(x_2)}(u_4)\right),$$

$$x_5 = F^{-1}_{r_{54};x_4}\left(F^{-1}_{r_{52|4};F_{2|4}(x_2)}(u_5)\right).$$

Conditional distributions $F_{2|3}(x_2), F_{2|4}(x_2)$ are not known and must be found as follows

$$\begin{aligned} F_{2|3}(x_2) &= \int F_{r_{12};x_1}(x_2) F_{r_{13};x_3}(x_1)dx_1 \\ F_{2|4}(x_2) &= \int c_{42|3}\left(F_{r_{43};x_3}(x_4), \int F_{r_{12};x_1}(x_2)F_{r_{13};x_3}(x_1)dx_1\right)dx_3 \end{aligned}$$

where $c_{42|3}$ is a density of the copula with correlation $r_{42|3}$.

Using diagonal band copula, for which conditional distributions are simple the above integrals can be calculated quickly but for bbns with many nodes calculations will be time consuming. Nonetheless, good numerical results are obtained in the sense that the specified conditional rank correlations are realized.

The main use of bbns is in situations that require statistical inference. After observing some events, we might want to infer the probabilities of other unobserved events. Alternatively, we might consider performing (expensive) observations and may wish to see how the results would impact our beliefs in other events of interest. In discrete bbns inference is complicated (Lauritzen S.L. ans Spiegelhalter D.J. 1998).
For continuous bbns updating is simple. Since the correlation $r_{ij|k}$ is equal r_{ij} if X_k is constant then if the given bbn can be represented as one D-vine, then updating in such a bbns requires only removing all edges incoming and ongoing to given node and dropping this variable from all (conditional) correlations on a vine. In case the bbn cannot be represented as one vine the value of the observed variable has to be used in sampling algorithm.

4 Example: Flight crew alertness model

In Figure 3 a flight crew alertness model adpated from the discrete model described in (Roelen A.L.C. et. al. 2003) is presented. In the original model all chance nodes were discretized to take one of two values "OK" or "Not OK". Alertness is measured by performance on a simple tracking test. The results are scored on an increasing scale and can be modelled as a continuous variable. Continuous distributions for each node must be gathered from existing data or expert judgement (Cooke R.M. 1991). The distribution functions are used to transform each variable to uniform on the interval (0,1). Required (conditional) rank correlations are found using the protocol described in Section 3. These can be assessed by experts in the way described in (Kraan B.C.P. 2002). In Figure 3 (conditional) rank correlation is assigned to each arc of the bbn. These numbers are chosen to illustrate this approach and are based on the auathors' subjective opinion.

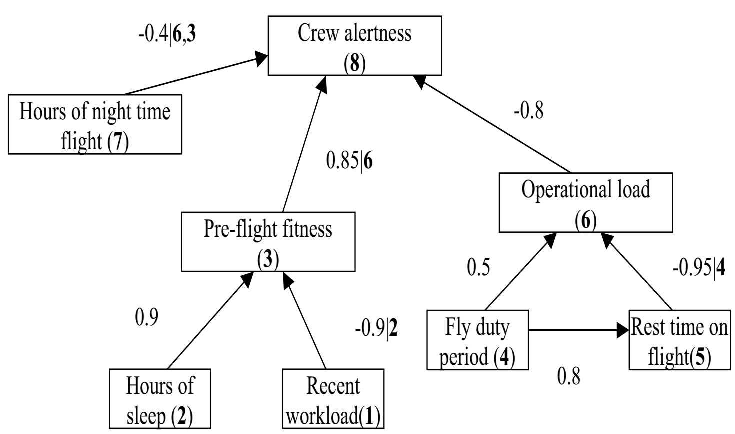


Figure 3: Flight crew alertness model.

The sampling algorithm for distribution described by the bbn in Figure 3 is the following.

$$x_1 = u_1,$$

$$x_2 = u_2,$$

$$x_3 = F^{-1}_{r_{32};x_2}(F^{-1}_{r_{13|2};x_1}(u_3)),$$

$$x_4 = u_4,$$

$$x_5 = F^{-1}_{r_{54};x_4}(u_5),$$

$$x_6 = F^{-1}_{r_{64};x_4}(F^{-1}_{r_{65|4};F_{5|4}(x_5)}(u_6)),$$

$$x_7 = u_7,$$

$$x_8 = F^{-1}_{r_{86};x_6}\left(F^{-1}_{r_{83|6|4};x_3}\left(F^{-1}_{r_{87|36};x_7}(u_8)\right)\right).$$

Using the above algorithm we obtained 10000 samples. In simulations the diagonal band copula was used. The correlation matrix corresponding to the bbn in Figure 3 is the following.

$$\begin{bmatrix} 1 & -0.0056 & -0.3788 & -0.0101 & -0.0102 & -0.0074 & -0.0095 & -0.1576 \\ & 1 & \underline{0.9014} & -0.0201 & -0.0193 & -0.0076 & -0.0027 & 0.4516 \\ & & 1 & -0.0156 & -0.0151 & -0.0060 & -0.0003 & 0.4927 \\ & & & 1 & \underline{0.8012} & \underline{0.4913} & -0.0072 & -0.4050 \\ & & & & 1 & -0.0757 & -0.0114 & 0.0430 \\ & & & & & 1 & 0.0026 & \underline{-0.7998} \\ & & & & & & 1 & -0.1179 \\ & & & & & & & 1 \end{bmatrix}$$

Notice that correlations in the matrix that correspond to rank correlations specified on the bbn (underlined) are correct up to the sampling error.

The main use of bbns is in decision support, and in particular updating on the basis of possible observations. Let us suppose that we know before the flight that the crew didn't have enough sleep and they will have along flight. Let us assume that the crew's hours of sleep correspond to 25th percentile of hours of sleep distribution and the fly duty period is equal to 80th percentile of fly duty period distribution.

We seek policies that could compensate loss of the crew alertness in this situation. Firstly we require that the number of night hours on the flight should be small (equal to 10th percentile). This improves situation a bit (dotted line in Figure 4). Alternatively we could require having long resting time on a flight (equal to 90th percentile). This results in a significant improvement of the crew alertness distribution (see dashed line in Figure 4). Combining these both polices improves the result even more.

Notice that in comparing different polices it is not necessary to know actual distributions of given variables. Our decisions can be based on quantile information. We might think of the transformation from quantiles to physical units of the variables as being absorbed into a monotonic utility function. Thus, conclusions based on quantiles will hold for all monotonic utility functions of the random variables.

Notice also that this quantification requires eight numbers. If the individual nodes are described with discrete distributions involving K outcomes, then $22K$ algebraically independent numbers are required. This demonstrates the dramatic reduction of assessment burden obtained by quantifying influence as conditional rank correlation.

5 Conclusions

The discrete bbns have recently become a very popular tool in modelling of risk and reliability. Their popularity is based on the fact that the influence diagrams capture engineer's intuitive understanding of complex systems, and at the same time serve as user interfaces for sophisticated software systems. Distribution free continuous bbns can significantly reduce the assessment burden, while preserving the interpretation arrows as influence.

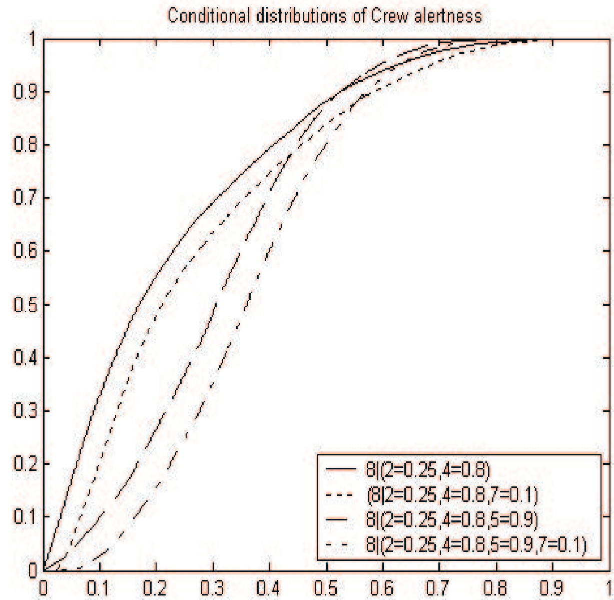


Figure 4: Four conditional Distributions of Crew alertness.

References

Bedford T.J., Cooke R.M. (2002). Vines - a new graphical model for dependent random variables. *Ann. of Stat. 30*(4), 1031–1068.

Cooke R.M. (1991). *Experts in Uncertainty*. New York: Oxford University Press.

Cooke R.M. (1997). Markov and entropy properties of tree and vines- dependent variables. In *Proceedings of the ASA Section of Bayesian Statistical Science*.

Cooke R.M. and Waij R. (1986). Monte carlo sampling for generalized knowledge dependence with application to human reliability. *Risk Analysis 6*, 335–343.

Doruet Mari D., Kotz S. (2001). *Correlation and Dependence*. London: Imperial College Press.

Joe H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Kraan B.C.P. (2002). *Probabilistic Inversion in Uncertainty Analysus and related topics*. Isbn 90-9015710-7, TU Delft.

Kurowicka D. and Cooke R.M. (2002). The vine copula method for representing high dimensional dependent distributions; application to continuous belief nets. *Proc. Winter Simulation Conference*.

Kurowicka D, Misiewicz J. and Cooke R.M. (2000). Elliptical copulae. *Proc of the International Conference on Monte Carlo Simulation - Monte Carlo*, 209–214.

Lauritzen S.L. ans Spiegelhalter D.J. (1998). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society 50*(B), 157–224.

Nelsen R.B. (1999). *An Introduction to Copulas*. New York: Springer.

Roelen A.L.C. et. al. (2003). Casual modeling for integrated safety at airport. *Proc ESREL 2003, Safety and Reliability 2*, 1321–1327.